Training and Testing on different distributions.

Case: We have 200,000 data from webpages as training set, and 10,000 data from mobile app which are from different distribution from training data. The 10,000 data are more similar to actual inputs for our model. Thus they are used as dev / test data.

Now we have the training and testing on different distributions problem.

Option 1: combine 10,000 mobile data with 200,000 web data.

Shuffle the new 210,000 dataset,

Randomly split 205,000 ⟶ training

2,500 ⟶ dev

2,500 ⟶ test

This is now a good idea, since now our dev and test sets most likely contain a lot more webpage images than mobile images, shifting them away from the actual input distribution. They wont be able to effectively evaluate our model.

Optition2 (better solution):

Split 10,000 mobile images to 5,000 / 2,500 / 2,500 groups.

add 5,000 to training set and

set 2,500 groups as test and dev sets.

Bias and Variance with mismatched data distributions

Training-dev: same distribution as training set, but not used for training.

Human Level	↕ avoidable bias
Training set error	↕ variance
Training-dev set	↕ data mismatched
Dev error	↕ degree of overfit to dev set
Test error	

Addressing data mismatch:

- Carrying out manual error analysis, understand differences between training and dev/test set.
- make training data more similar, or collect more data similar to dev/test sets.

one approach: artificial data synthesis

e.g. human voice audio + car noise = synthesized in-car audio